DOCUMENT RESUME

ED 247 267 TM 840 440

AUTHOR Pearlman, Mari Ann
TITLE Theory and Practice: The Revised Joint Technical Standards and Test Construction.
PUB DATE Apr 84
NOTE 17p.; Paper presented at the Joint Annual Meetings of the American Educational Research Association and the National Council on Measurement in Education (New Orleans, LA, April 23-27, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Viewpoints (120)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Entrance Examinations; Racial Discrimination; Sex Discrimination; *Standards; *Test Bias; *Test Construction; Test Items
IDENTIFIERS *Standards for Educational and Psychological Tests; *Test Content

ABSTRACT

Two of the standards contained in the third draft of the new Joint Technical Standards for Test Development and Revision are discussed: Standard 3.13, mandating the use of multicultural material and the avoidance of material offensive to any major ethnic, cultural, or gender group; and Standard 3.14, mandating research and subsequent test revision to eliminate aspects of test design, content, or format that might serve to bias test scores positively or negatively for any given group. In evaluating these standards, test developers must keep in mind the purpose of the test being developed. They should also realize that these two standards imply that including multicultural material will ensure that major subgroups will see material familiar to them and thus score better on the test (an assumption that is unproven). Research addressing this assumption must study items with two characteristics: (1) that differential performance has been detected on these items, and (2) that the context of the items can be changed to a context relevant to subgroup culture without altering the essential task. This is almost impossible. The fourth draft of the Standards revised these two standards. Standard 3.13 (now 3.5) was made more general and less prescriptive; and the two standards were separated. (BW)

Theory and Practice: The Revised Joint Technical Standards and Test Construction*

Mari Ann Pearlman

Educational Testing Service

This paper is based on the third draft of the Joint Technical Standards, the latest draft available at the time of writing. In the third draft of the new Joint Technical Standards for Test Development and Revision, two items stand out as of particular interest and concern to test developers, and it is upon those standards that I will focus my remarks. They are Standard 3.13, which mandates the use of multicultural material and the avoidance of material offensive to any major ethnic, cultural, or gender group and Standard 3.14, which mandates research and subsequent test revision to eliminate aspects of test design, content, or format that might serve to bias test scores positively or negatively for any given group.

---

* Symposium paper presented at the Annual Meeting of the American Educational Research Association and the National Council on Measurement in Education, New Orleans, April 1984.

No test developer would, I believe, oppose the philosophical position these standards embody. A test whose content accurately reflects, to quote the language of Standard 3.13, "the cultural backgrounds and prior experiences of the major ethnic, cultural, and gender groups represented in the intended population of test takers" is likely to be not only broader in its scope but also much more interesting to develop and take. And a test whose items do not provide undue differential advantages or disadvantages to specific groups of test takers should clearly be a desideratum for all test development.

In fact, perhaps the most important challenge facing professional test developers in the next 20 years is the development of measurement instruments that are not biased toward any test taker. At present, however, I believe that we must be aware of the technical limitations of the state of our art. We must not delude ourselves on such an important issue; the modesty of our accomplishments in this area thus far must be acknowledged, and the reasons for that modesty explored. Attempting to put these standards to work in the daily business of writing items and compiling tests reveals difficulties of intepretation. In evaluating Standards 3.13 and 3.14 we must keep in mind a very simple question: What is the purpose of the test we are developing? It has a very limited, practical function: it is a device to measure certain narrowly defined skills. To perform

this function adequately, each form of the test must be parallel with previous forms, and scores on any one form must correlate at least as well as they have in the past with whatever external standard is used to measure validity. In the case of national admissions tests, for example, this standard is the grade point average of freshmen in college. Therefore, the general framework within which we consider Standards 3.13 and 3.14 is determined by awareness of just what flexibility and latitude we have within these predetermined constraints.

Turning first to Standard 3.13, the inclusion of multicultural materials, we are confronted immediately with certain practical problems. In national admissions testing, the intended population of test takers can include just about anyone. How are we to define a "major" cultural and/or ethnic group? Is our responsibility discharged if only protected minorities are addressed? How much reflection of diversity of background, and whose diversity, do we put into the test? Because multicultural material must be inoffensive to everyone, it may be impossible to consistently reflect the cultural backgrounds and prior experiences of major cultural and ethnic subgroups in the population of the United States, or to reflect accurately the prior experiences of women. Those interactions between the subgroups and the majority culture that were most significant in the prior experience of the subgroups are very likely to be in some way offensive if accurately explored in texts of sufficient complexity to meet the goals of this Standard.

The Comment appended to Standard 3.13 suggests the establishment of a review process for all materials to detect and eliminate material likely to be offensive to groups in the test-taking population. This can be done, and indeed is done routinely at Educational Testing Service, where we have developed a special Test Sensitivity Review procedure that is obligatory for all tests. The following are some of the standards used by trained sensitivity reviewers in reviewing test materials. [Copy of ETS Sensitivity Guidelines handed out to audience]

While the application of these standards certainly eliminates very obvious and offensive stereotypical language from test content, an accomplishment not to be lightly dismissed, it does not even attempt to truly "reflect the cultural background and prior experiences" of major subgroups. What happens in actual practice is that the language of all materials is very carefully scrutinized; women do appear by feminine pronouns in mathematics items, men are not always the movers, shakers, thinkers, and authors of all. Clearly, Standard 3.13 intends to engender more searching efforts than these on the part of test developers. And here we must return to the general framework I suggested before: to what end should we seek to include materials that really "reflect the cultural background and prior experiences" of major subgroups? Taken seriously, this could mean including materials that possibly only a small group of test takers could identify with, thus introducing a new bias into the test. Why does Standard 3.13 mandate such attempts at "fairness?"

A partial answer to that question is implied, I believe, in the spaces between Standards 3.13 and 3.14. In a second point made in the comment appended to Standard 3.13 it is argued that a review process like the one described above is no substitute for attention to the different cultural-experiential bases relevant to test material in the item-construction stage and, before that, in the test and domain specification stage of test development.

This comment implicitly ties Standard 3.13 to 3.14, which concerns itself with differential performance on test items. Such a connection needs to be very carefully scrutinized. It is by no means clear that the inclusion of multicultural materials will in and of itself have any effect on the differential performance of subgroups on specific test items. The inclusion of materials that attempt to broaden the subject matter base of the test and to avoid perpetuating stereotypical and biased ways of thinking about ethnic and cultural subgroups is worth doing in and of itself, because it is intellectually honest and responsible. But to suggest that such a procedure bear the burden of reducing or eliminating differential performance is unrealistic as well as empirically suspect. It is not clear whether material actually known to be offensive to major subgroups in the test-taking population would differentially affect performance, for it is a hypothesis virtually impossible to test responsibly. One assumes that such material would prejudice performance, but we do not test this hypothesis for reasons analogous to those used by laboratory

scientists who do not test the effects of massive doses of suspected carcinogens on human subjects. Furthermore, there is no firm information or broad agreement about what specifics of the cultural backgrounds and prior experiences of various subgroups affect performance on standardized tests nor how they do whatever affecting they may do. Beyond very general characteristics that affect performance on standardized tests, like socioeconomic status and amount and breadth of schooling, we can not identify other differences, if there are any, that create an intellectual problem-solving style unique to a particular subgroup, a style that would affect performance on standardized tests.

The implicit assumption in these two Standards is that including multicultural material would automatically ensure that major subgroups would see material familiar to them and thus score better on the test. A moment's thought will convince us that this is first, an unproven assumption and second, an unworkable suggestion. Many constraints govern the content of any form of a test--in reading comprehension, the only place in which passages long enough to deal with these subjects appear, concern for the differing experiences and interests of students mandates inclusion of material from a variety of areas, such as natural science, social science, humanities. Futhermore, any one test form will include, at most, five or six reading passages distributed among these areas of interest. Because each test taker sees only one form of a test, he or she is unlikely to encounter a passage that

reflects his or her cultural background and prior experiences. Thus, satisfying the standard becomes an aesthetic achievement for a testing corporation; all the diverse subgroups are represented over a series of test forms. Such a procedure has virtually no impact at all on the test takers themselves.

The materials now included in national admissions tests conform to a certain basic model. This model helps define the set of important skills, verbal and quantitative, students need to be successful, that is, to get good grades, in college. Substantial modification of the content of the materials should be based on explicit rationales and justifications.

With all these reservations made clear, I should like to examine the assumptions of Standard 3.14 in light of my experience on just such a research project as the Standard suggests is desirable. The first step in such a study is the detection of item bias, the least difficult part of the research, and itself a vexed issue. At present test developers and researchers are far from unanimity on the best statistical model to use for detecting item bias. And of course different items will be identified as biased depending upon which statistical model is used. Once some method has been chosen and biased items are identified, even greater difficulties and ambiguities arise in an attempt to formulate hypotheses that might explain the bias: what is it about these particular items that causes different groups of test-takers to perform unusually well or unusually poorly in

comparison to their performance on the total section or test? It comes as no surprise, I'm sure, that generating hypotheses to explain the relatively poor performances of major cultural and ethnic subgroups on certain test items is not very difficult. We think we know, or at least suspect, what characteristics of test design and item format might produce disadvantages for these test-takers. It is much more challenging, however, to hypothesize what it is about a group of items that puts the higher scoring group, such as men on quantitative items, at a disadvantage. And hypothesizing about what makes for performance substantially above the expected level on certain items among subgroups of test takers is equally problematic. In general, the items look pretty similar. They have been developed using the same guidelines and format, the range of difficulty of the items as revealed in pretesting doesn't explain much about differential performance, and the subject matter seems to have little bearing in most cases on performance.

However, even though hypothesis formulation is fraught with problems, once it is completed the clear-cut parts of such research are over. For now, items similar to those identified as biased must be revised in order to test the validity of the hypothesis. Note that the same items are not typically revised. Thus, if the bias arose originally because of some quality peculiar to an item or set and not reproduced in another, all hypotheses are confounded. Also, the process of revision, at least in verbal items, introduces so many confounding variables

that interpretations of the results must be very carefully hedged and limited. Let us say, for example, that one hypothesis is, as it was in the study in which I participated, that Reading Comprehension questions which have stems using the LEAST, NOT, or EXCEPT format, like this one [1] are likely to bias performance against Black test takers. We hypothesized, for the purposes of the study, that Black test takers might be at a substantial disadvantage in performing such a task, which asks not for the one right response, but for the one anomalous, different, or wrong response. In revising such an item to test this hypothesis, the stem was altered to read like this [2]. However, because we are now asking for the one right response, we changed at least three options, thus essentially creating a new question. Even if the results indicate the expected change in performance, I think it unlikely that we could say with any degree of certainty that such results substantially increase the probability of our hypothesis.

But what we really want to find out by such research is even harder to discover. We would like to test the implicit connection made in these Standards between multicultural materials and differential performance. To examine the accuracy of this hypothesis, that a particular subgroup will feel more confident dealing with and thus do better on a task set in a context familiar to it, the researcher must select items with the following characteristics: 1) differential performance has been detected on these items, and 2) the context of the items can be changed to a context relevant to subgroup culture and experience

without altering the essential task. This is a tall order. Context and essence seem to be so interwoven as to be inseparable in items involving reading and comprehension. In certain obvious cases content causes item bias. An analogy which uses terms such as "biretta" like this one [3] clearly favors those test takers with a Roman Catholic background. Ironically, of course, this might include a substantial proportion of another major minority subgroup, like Hispanic Americans. Clearly, too, items like these [4] favor a small segment of the test taking population and should be avoided. But the more fundamental questions about biases built into test designs and item format and content remain very difficult to get at in an organized empirical fashion.

In the fourth draft of the Standards, which became available in March after this paper was written, the two standards I have here discussed have been revised, one of them substantially. I applaud the revision of the Standard concerning the inclusion of multicultural materials (originally 3.13, now 3.5), which has been made much mere general and less prescriptive. Also, the two standards have been separated in the chapter on Test Development, a wise revision, given the implications of putting them back to back, which I have discussed. Any return to the specificity of the Third Draft Standards would be a serious misjudgment of the task of test developers in my view.

All of the following are used by the author in making his point EXCEPT

(A) a chronological analysis
(B) a specific illustration
(C) an explanation of a term
(D) a statistical generalization
(E) an appeal to authority

The author uses which of the following techniques in making his point?

(A) A chronological analysis
(B) A personal narrative
(C) A formal definition
(D) A tabulation of statistics
(E) An appeal to authority

1/2

All of the following describe the process by which melanin is formed EXCEPT:

(A) It is genetically controlled.
(B) It is a sequence of reactions.
(C) It can take place in different kinds of cells.
(D) It begins with the formation of a colorless substance.
(E) It requires a certain enzyme for completion.

Which of the following describes the process by which melanin is formed?

(A) It is not genetically controlled.
(B) It is a single repeated reaction.
(C) It can take place in different kinds of cells.
(D) It begins with the formation of a pigmented substance.
(E) It requires a certain enzyme for completion.

MORTARBOARD : ACADEMIC : :

(A) turban : monastic

(B) cap : youthful

(C) wimple : classical

(D) biretta : ecclesiastical

(E) helmet : medieval

L A C R O S S E : S T I C K : :

(A) b o x i n g : g l o v e

(B) s w i m m i n g : w a t e r

(C) t e n n i s : n e t

(D) s q u a s h : r a c k e t

(E) b a s k e t b a l l : g o a l

O P E R A : A R I A : :

(A) b a l l e t : p i r o u e t t e

(B) p l a y : s o l i l o q u y

(C) p o r t r a i t : c a n v a s

(D) o r c h e s t r a : m a e s t r o

(E) c o n c e r t : s o l o i s t

Presentation II: Theory and Practice: The Revised Joint Technical Standards and Test Construction

Mari Pearlman
Educational Testing Service

This presentation will discuss aspects of the revised Joint Technical Standards as they affect the ongoing process of test construction for national admissions testing.

Perhaps the most farreaching assumptions in the revised Joint Technical Standards from this perspective are those that imply the desiderata of the training and background of test developers. These tacit assumptions and their implications for policies and procedures will be discussed. Two specific parts of the standards, those mandating the consideration of test material from a multi-cultural, ethnic- and gender-sensitive viewpoint, and those directing test developers to study differential performance on test items, will be examined, with specific examples of problems and solutions in these areas presented. Finally, some social and policy implications of both the assumptions and the requirements of the revised Joint Technical Standards will be addressed.